

BAB 1. PENDAHULUAN

1.1. Latar Belakang

Penggunaan media sosial di Indonesia telah meningkat pesat selama beberapa tahun terakhir yang berperan sebagai ruang publik utama guna pertukaran informasi, opini, dan mobilisasi sosial. Platform mikroblogging seperti X dan Threads memungkinkan penyebaran opini secara cepat dan luas, namun sekaligus menciptakan tantangan besar terkait moderasi konten karena tingginya volume serta sifat pesan yang singkat dan mudah diselewengkan. Dampak dari penyebaran konten negatif termasuk ujaran kebencian, perundungan siber, dan ajakan yang melanggar aturan hukum tidak hanya bersifat psikologis bagi individu tetapi juga berpotensi memicu konflik sosial dan masalah hukum yang lebih luas. Dalam konteks ini, upaya otomatisasi deteksi konten bermuatan kebencian menjadi kebutuhan penting bagi penegak kebijakan, penyedia platform, dan peneliti (Dhafa Maulana et al., 2025).

Dalam kajian komunikasi dan ilmu sosial, istilah diskursus dipahami bukan sekadar sebagai pertukaran bahasa atau opini, melainkan sebagai sistem pengetahuan yang membentuk cara berpikir, berbicara, dan memaknai realitas sosial. Mengacu pada pemikiran Michel Foucault, diskursus merupakan kumpulan pernyataan, ide, dan praktik bahasa yang terorganisasi serta beroperasi dalam relasi pengetahuan dan kekuasaan tertentu, sehingga menentukan apa yang dapat dikatakan, dipikirkan, dan dianggap benar dalam suatu konteks sosial (Kurniawan, 2023). Dengan demikian, diskursus tidak hanya merefleksikan pandangan individu, tetapi juga membangun kerangka makna kolektif yang memengaruhi persepsi masyarakat terhadap suatu isu, kebijakan, maupun kelompok sosial. Dalam konteks media sosial, diskursus berkembang secara dinamis melalui interaksi antarpengguna yang dipengaruhi oleh emosi, ideologi, serta kepentingan institusional sehingga menjadikannya kompleks, kontekstual, dan rentan melahirkan praktik bahasa bermuatan kekuasaan, termasuk ujaran kebencian yang

kerap tersamarkan dalam bentuk kritik, satire, atau ekspresi emosional. Perdebatan publik tentang Rancangan Kitab Undang-Undang Hukum Acara Pidana (RKUHAP) memicu lonjakan diskursus politik di media sosial yang kerap mengandung ujaran kebencian, kritik tajam, dan bahasa emosional. Kondisi ini membuat pemisahan antara kritik yang sah dan ujaran kebencian menjadi sulit dilakukan secara otomatis, terutama ketika ekspresi kebencian disampaikan secara implisit melalui sindiran, sarkasme, atau bahasa emosional.

Berdasarkan observasi awal terhadap diskursus RKUHAP di media sosial, pada periode bulan November - Desember 2025 publik menunjukkan tingkat aktivitas yang cukup tinggi pada platform X dan Threads, ditandai oleh intensitas unggahan, penggunaan hashtag, serta interaksi antar pengguna. Dalam diskursus tersebut, selain kritik terhadap kebijakan, ditemukan pula penggunaan bahasa emosional yang berpotensi mengandung ujaran kebencian. Untuk memberikan gambaran empiris awal mengenai kondisi tersebut, dilakukan observasi kuantitatif awal terhadap 2.385 data komentar, yang terdiri atas 183 data ujaran kebencian dan 2.202 data non-ujaran kebencian, sebagaimana dirangkum pada Tabel 1.1.

Tabel 1. 1 Observasi Awal Diskursus RKUHAP di Media Sosial

Platform	Tagar	Non ujaran kebencian	Ujaran kebencian
X	#SemuaBisaKena #tolakruukuhap	1.955	164
Threads	#SemuaBisaKena #tolakruukuhap	247	19

Hasil observasi awal tersebut menunjukkan bahwa diskursus RKUHAP di media sosial memiliki potensi kemunculan ujaran kebencian sebagai bagian dari dinamika interaksi publik. Temuan ini menjadi dasar perlunya pendekatan otomatis berbasis pemrosesan bahasa alami untuk mengidentifikasi ujaran kebencian secara lebih sistematis dan akurat.

Penelitian-penelitian sebelumnya pada kasus serupa menunjukkan variasi besar pada pendekatan dan performa model model tradisional seperti *K-Nearest*

Neighbors (KNN) yang menggunakan TF-IDF menunjukkan kegunaan dasar namun terbatas dalam menangkap nuansa dalam diskursus sosio-legislatif (Septiawan et al., 2025), sedangkan pendekatan *deep-learning* berbasis *transformer* seperti IndoBERT umumnya memberikan representasi kontekstual yang lebih kaya dan performa jauh lebih baik pada tugas klasifikasi berbahasa Indonesia (Dharmawan et al., 2023).

Secara lebih rinci, berbagai penelitian menunjukkan bahwa model yang memanfaatkan IndoBERT dalam arsitektur klasifikasi memiliki kemampuan yang lebih baik dalam memahami konteks Bahasa Indonesia yang kompleks, termasuk variasi bahasa sehari-hari, penggunaan istilah lokal, serta struktur kalimat yang tidak baku. Hal ini membuat IndoBERT sering kali unggul dibandingkan model berbasis urutan klasik dalam tugas deteksi ujaran kebencian maupun pelanggaran yang berkaitan dengan UU ITE. Sebaliknya, pendekatan konvensional yang mengandalkan *pipeline* seperti TF-IDF dan KNN cenderung menghasilkan performa yang lebih terbatas pada korpus text bertema RKUHP. Model tersebut dinilai kesulitan dalam menangkap makna kontekstual dari teks, terutama ketika menghadapi fenomena seperti sarkasme, ambiguitas makna, serta penggunaan istilah yang bersifat kiasan, sehingga kinerjanya relatif moderat. Di sisi lain, penerapan jaringan saraf sederhana seperti feedforward neural network yang dikombinasikan dengan embedding IndoBERT memang mampu mencapai tingkat akurasi yang cukup tinggi pada dataset tertentu, namun masih menunjukkan kelemahan ketika harus membedakan antara ujaran kasar dan ujaran yang benar-benar mengandung unsur kebencian, khususnya ketika konteks percakapan tidak sepenuhnya dipahami.

Temuan-temuan tersebut mengindikasikan adanya dua poin penting yang menjadi landasan utama penelitian ini. Pertama, kekuatan representasi kontekstual dari IndoBERT menjadikannya pilihan yang menjanjikan untuk digunakan dalam mendeteksi ujaran kebencian pada teks berbahasa Indonesia di media sosial x dan threads oleh model IndoBERT, terutama dalam diskursus sensitif seperti pembahasan RKUHAP yang sarat dengan opini, emosi, dan muatan politis. Kedua, kondisi nyata data di media sosial menampilkan berbagai tantangan yang tidak bisa

diabaikan, seperti ketidakseimbangan jumlah data antar kelas, kemunculan bahasa sarkastik, slang, serta pergantian kode, hingga kecenderungan model untuk melakukan kesalahan klasifikasi (false positive) ketika hanya mengandalkan kemunculan kata-kata kasar tanpa memahami konteks secara menyeluruh. Oleh karena itu, penelitian ini difokuskan pada penerapan dan pengujian model IndoBERT yang di fine-tune secara khusus pada korpus diskursus RKUHAP di media sosial X dan Threads, dengan harapan mampu menghasilkan model yang lebih adaptif, kontekstual, dan akurat dalam membedakan antara kritik, opini, dan ujaran kebencian yang sebenarnya.

1.2. Rumusan Masalah

Berdasarkan latar belakang yang sudah di uraikan, maka rumusan masalah dalam penelitian ini adalah sebagai berikut :

- a. Bagaimana pengelompokan ujaran kebencian yang muncul dalam diskursus mengenai RKUHAP di media sosial X dan Threads?
- b. Bagaimana proses penerapan model IndoBERT dalam mendeteksi ujaran kebencian pada unggahan yang berkaitan dengan Rancangan Kitab Undang-Undang Hukum Acara Pidana (RKUHAP) di media sosial X dan Threads?
- c. Seberapa baik performa model IndoBERT dalam mengklasifikasi ujaran kebencian?

1.3. Tujuan Penelitian

Adapun tujuan dari penelitian ini adalah sebagai berikut :

- a. Untuk mengidentifikasi dan menganalisis pengelompokan ujaran kebencian dalam diskursus RKUHAP di media sosial X dan Threads.
- b. Untuk mengimplementasikan model IndoBERT dalam mendeteksi dan mengklasifikasikan ujaran kebencian pada text yang berkaitan dengan RKUHAP.
- c. Untuk mengukur dan mengevaluasi kinerja model IndoBERT dalam mendeteksi ujaran kebencian.

1.4. Manfaat Penelitian

Manfaat yang diharapkan dalam penelitian ini meliputi :

1. Membantu upaya pencegahan dan pengawasan penyebaran ujaran kebencian dimesia sosial X dan Threads, khususnya yang berkaitan dengan isu sensitif seperti RKUHAP.
2. Menambah referensi akademik mengenai penerapan model IndoBERT dalam mendeteksi ujaran kebencian dalam diskursus hukum dan isu politik yang ada di media sosial.
3. Memperkaya kajian penelitian sebelumnya yang masih terbatas pada penggunaan model tradisional seperti KNN atau Bi-LSTM dalam kasus ujaran kebencian

1.5. Batasan Masalah

Agar penelitian ini lebih terarah dan fokus pada tujuan yang telah ditetapkan, maka batasan masalah dalam penelitian ini adalah sebagai berikut:

1. Populasi data yang digunakan berjumlah 6.450 data teks bersih dari media sosial X dan Threads dan Data dibagi menggunakan rasio 80% data latih sebanyak 5.160 data, 10% data validasi sebanyak 645 data, dan 10% data uji sebanyak 645 data.
2. Menggunakan model dasar *IndoBERT-Base-p1*